NWX-US DEPT OF COMMERCE

Moderator: Kim Brown
March 13, 2014
2:00 pm CT

Coordinator: Welcome and thank you for standing by. At this time all participants' lines will remain on a listen-only mode until the question and answer session.

During the question and answer session please press Star 1 on your touch-tone telephone. You will be prompted to record your name in order to be introduced.

Today's conference is being recorded. If you have any objections you may disconnect at this time.

I would like to turn today's call over to Tim Gilbert, American Community Survey. Sir you may begin.

Tim Gilbert: Thank you very much. My name is Tim Gilbert. I'm from the US Census Bureau. And today I'll be talking about the ACS Public Use Micro Data samples or the ACS PUMS files as we refer to them and specially talking about the multiple vintage variables on our 2012 multi-year files.

So I'll start out talking - going over quickly some of the fundamentals of PUMS data before we get right into discussing the multi or dual vintage variables.

And one of the basic fundamentals of PUMS is understanding the difference between summary data and micro data. And the summary data as a lot of people are familiar with are pre-tabulated formatted tables that we produce and release mainly through American FactFinder.

And you can see an example of a summary table up on the top left of this slide. And they're pre-defined tables for specific geographic areas whereas the PUMS is micro data.

And as you can see on the bottom right a screenshot of an example of a micro data file showing that each row is representing an individual person or we also have housing unit files as well.

So they're set up so that you can create your own tables with the micro data.

And so basically as I mentioned PUMS it's the Public Use Micro Data Sample. And it's public use because it's been collected and then fully anonymized and it's made downloadable and accessible to the public.

And it's micro data as I referred to being that there are records of individual people or individual housing units.

And it's also a sample meaning that it's a representative of the population as a whole for whatever geographic area it's available in and that you're looking at.

And so the PUMS is also a subsample of all of the ACS interviews. So as the full ACS sample that we use to create the pre-tabulated tables in the summary data we take a subsample of those interviews, anonymize those and put those out as public use micro data sample.

And each one year of PUMS represents about 1% of all US households.

And the PUMS is also a weighted sample meaning that you'll need to use the weighting variables that we provide in the files to be able to create estimates and also calculate standard errors and margins of error.

And the PUMS comes in a set of two files, one for housing units and one for persons. You can use each individually or we have information about how to combine those files if you wanted to use variables off of each of the housing and the person's file to create estimates.

And we produce a one, three and a five year PUMS files very similar to the one, three and five year estimates that we release from the full ACS sample.

And we provide those PUMS files in both SAS format as well as CSV format. And you can also access the PUMS via DataFerrett which is a data tool we'll talk about a little bit later. But it's online tabulation tool where you can use the PUMS data and create your own tables online.

So some of the reasons for why someone might want to use PUMS rather than our summary data is that if you - if data are needed for a tabulation or a specific universe that we don't have available in the pre-tabulated tables such as if you wanted to look at single year of age which we don't provide in the pre-tableted summary tables.

Also if you were doing any kind of correlation or regression analysis or any kind of statistical analysis looking at relationships between different characteristics or different variables you can use the PUMS file to do that.

And you can also create new measures that are not available in our pre-tabulated tables using multiple different variables. And you can combine those and look at small subsets of the population.

So as I mentioned we released a one year, a three year and a five year PUMS file every year. And our most recent one year for the 2012 data year was just released in December or 2013.

And then we released our three year, 2012 three year PUMS in February. And just last week we released our five year PUMS in March this month.

And as you - if you started working with these multi-year, these 2012 multi-year files you might notice or you might already know that we do have multi-vintage variables which is the point of this Webinar today on both of our three year and our five year files.

And the multi-year files contain all of the same cases in their component one year files. And so we bring these together and then we update the weights that we use to represent the latest vintage of them, the most recent year as well as we - as standardizing the dollar amounts.

We standardize those across all the years so that you can get the most recent inflation adjusted dollar amounts.

And some of the reasons for using the multi-year PUMS files would be for studying small groups where you need a larger sample to analyze these very

small subpopulations or if you're already conducting analysis with some of our pre-tabulated summary data that is using multi-year and you wanted to compliment that with the PUMS then we have multi-year files with the PUMS as well for the same time spans.

So now we'll start talking a little bit about the multiple vintage variables that we have in our PUMS files and our multi-year PUMS files for 2012.

And occasionally we have to update some of the response categories or some of the categories in our code list to reflect updated classification systems used for analysis of the economy or changes in detailed race, ancestry or place of birth codes or due to data disclosure requirements that come up.

And as we update these values that can cause the multi-year PUMS files to contain multiple vintages for different years within the same file.

And so essentially the multi-vintage variables are essentially the same variable across years except for certain individual years they might have differing sets of values for the different years within the same file.

And so here's a table showing all of our different variables in the 2012 three year and the 2012 five year PUMS files. And the different - these are all the variables that have multiple vintages as well as the variable names for each vintage.

So for example in the 2012 three year file where we used to have the variable for first ancestry ANC1P we now have two vintages of that variable.

And so for the years 2010 and 2011 we have the variable name ANC1P05. And for the 2012 year we have the variable ANC1P12.

And we also have in the 2012 five year PUMS files we have two variables with not just two vintages but three vintages. And those variables are the related to occupation.

And then the OCCP variables as well as the SOCP variables. So we have an 02 vintage which relates to the years 2008 and 2009, the ten - 2010 vintage which relates to the years 2010 and 2011 and then a 2012 vintage which relates to the last year and the five year file in the most year of 2012.

One of the other variables that is contained in the multiyear PUMS files that does have a dual vintage is the PUMA or Public Use Micro data Area variable. That's the - as you might be aware that's the lowest level of geography that's contained in the PUMS files.

And so now we'll talk a little bit about the two different vintages that we have in our PUMS files for PUMA.

And so some of the other geographic identifiers that we - that are contained in the file is the region, the division, the state level variable as well as the PUMA being the lowest level of geographic detail that we contain in the PUMS files.

And PUMAs can be used to identify geographic areas of at least 100,000 people. And so because of that and that the PUMA being the lowest level of geography contained in the PUMS files, the PUMS is not really designed for statistical analysis of very small geographic areas.

And as I mentioned we had two different vintages of the PUMS in our 2012 three year as well as 2012 five year files.

And the PUMAs are updated and redefined after each census. And they're defined by the states in coordination with our geography division here at the census bureau. And a lot of detail about how that process works can be found at the link displayed here on this Web page of I'm pointing you towards the Geography Division's Web page talking about the definition of the new PUMA boundaries based on the 2010 census.

And so we have - we're now including those definitions for the new 2010 boundaries in our 2010 files for the first time.

So you might have - we had those new definitions in our single year 2012 but now we're running into the issue of having multiple vintages for PUMAs in our 2012 three year files as well as the 2012 five year files.

And these - the PUMAs have to be large enough to meet disclosure avoidance requirements. And they're identified by a five digit number that's unique within each state.

So we have a variable on the file for the state as - that you can use in conjunction with the PUMA variable that contains the five digit unique number within each state.

And we've also created reference maps for both vintages of the PUMAs. You can access these reference maps via the Web page listed, the URL listed at the top of this slide.

And at this Web page you can expand the public use micro data areas section. And you can see that we have 2010 PUMA reference maps as well as Census 2000 PUMA maps which are the 5% sample that you see here.

And we also have a newer online interactive mapping product called Tiger Web that you can access from Geography's main page listed here on this slide.

And the Tiger Web decennial option contains boundaries that you can access on an interactive map for both the census 2000 vintage PUMAs which are used in years prior to the 2012 PUMS as well as the census 2010 vintage PUMAs which are used in the 2012 year of the multiyear PUMS.

So now that we've talked about some of the basics of the PUMS and what these multiple vintage variables are I'll actually get into a few examples of how to go through and how to think about working with the multiple vintage variables as well as the multiple vintage PUMAs.

And so some of the documentation that I'll be making use of in working with these multi-vintage variables on our PUMS files they can - all of these documents can be accessed from our PUMS documentation page which you might be familiar with. But we also have the URL listed here on this slide. And I'll be talking in more detail about the documentation page a little later.

So the first document that I'll be referencing is the PUMS readme which contains a as well as information about working with the PUMS files. We've also included a list of all the variables that have multiple vintages.

The PUMS data dictionary is the next document we'll be referencing. And that file contains all the variable names that are contained in both the housing and the population level PUMS files as well as the descriptions of those variables and the different values for each variable.

And then there's also the accuracy of the PUMS document is a very valuable document that can be referenced for information about working with PUMS files in general as well as some other examples that I won't be going through today that we have on working with multiple vintage variables as well as multiple vintage PUMAs.

So when working with a multiple vintage variable if I were interested in working with marital history or the year of last marriage and I had worked prior in previous years with the variable MARHYP I would first go to the PUMS readme document if I wanted to work in the 2012 five year PUMS file.

And I would verify to see if this variable I'm used to working with is actually a multiple vintage variable.

And in Section 7 talking about the variable changes we can see that MARHYP is now split into two variables, one with an 05 suffix and a 12 suffix.

So knowing that the next step that I would take would be to go to the data dictionary for the 2012 five year file and I would look up both of those variables to see exactly the values that differ between each vintage.

And so for MARHYP05 I see that the lowest value or the lowest year included in the value list is 1928. But with the '12, the 2012 vintage I see that that lowest year is 1932.

So right here we see the difference that we're working with.

And so the next step that I would take that I would know I need to take in working with these variables is I know that I would need to combine the two variables and standardize their value list to create a new derived variable.

And I would do that in this case by looking at if the MARHYP05 is less than or equal to 1932 because I know that's the lowest level that MARHYP 12 has.

So I'd look for less than or equal to 1932 for the '05 vintage. Or if the variable, the 2012 vintage equals 1932. Then I can create a derived variable that is - has a lowest value equal to 1932 that's standard across all five years that are included in the file.

So by doing this I know that I've created a variable that is going to have a value that's standards for every case or every record regardless of which year the data was collected in.

And so now that I've worked with a multiple vintage variable and created a new variable that's standard across all the year and the 2012 five year I know now that I'll be working with the Public Use Micro data area.

And I would first start by looking for knowing that we do have two vintages of the PUMA variable on the 2012 five year file. I would start by looking in the data dictionary to get the new PUMA variable names.

And I see that the two PUMA variables we now have are PUMA00 for the earlier 2000 based vintage and then the variable PUMA10 which relates to the newest 2010 PUMA vintage that's used in the 2012 year on the 2012 five year file.

So once again PUMA00 is going to apply to 2008, 2009, 2010 and 2011. And ten PUMA10 is going to apply to cases in the year 2012.

So the next step that I would take would then be making use of a very valuable resource at the Missouri State Data Center's Web site, their MABLE Geographic Correspondence Engine which you can access at the URL listed on this slide.

And if I'm interested in the state of Michigan I would first select Michigan from the state list. And then because I know that I've worked with - I know the 2000 PUMA base code that I've worked with before I would first select that I want to start with my source geo code would be the PUMA 2000 base. And then I would compare that to see the correspondence between that geographic area that I'm used to working with and the PUMA 2012 vintage to compare these two PUMAs.

And then I would get output showing me the correspondence between the PUMA 2000 based vintage and the PUMAs that are used in the 2012 year which are based on the 2010 census.

And if I'm working with - if I know that I've worked with PUMA 03806 before I can see that this has a direct one to one relationship that you can see by looking in the far right-hand column.

And now I can see that the new vintage PUMA code for the exact same geographic area is 03204.

So I know that much like the multi-vintage variable of MARHYP that we were working with earlier, I know that I have to create a sort of crosswalk to

be able to use both of these PUMA vintages and standardize them into a new derived variable that'll cover all five years.

And so to do that I would look at when if the state equals 26 which refers to the state of Michigan and PUMA00 equals 03806 which is the old vintage PUMA or if the state equals 26 for Michigan and the new vintage PUMA PUMA 10 equals 03204 because I know that those two PUMAs have a direct - directly correspond to each other than I can create a new derived variable for PUMA that I can give whatever value I want to.

And I can know that now these - this new derived PUMA variable is going to be standardized across all five years.

And so now I'll briefly go through accessing some of the PUMS data. And the first way to access the PUMS data would be to go into American FactFinder.

And you can get there by going to the PUMS Data page on the ACS from the ACS main page. When you go follow the URL link listed here you can click on PUMS Data.

And then you can if you're interested in the 28 through 2012 ACS five year PUMS you can click the five year PUMS link.

And this link will take you to American FactFinder where you can see that the public use migrated sample and the data set that you're - that we're interested in, the 2012 five year estimates are already listed in our selections. So that gives us options for downloading these files in either the CSV format or the SAS format.

And if you're not familiar with working with PUMS and statistical software packages such as SAS or SPSS, or Stata the Census Bureau has an on line data analysis and extraction tool called DataFerrett that you can access all of our PUMS data and create tables actually within this program.

And you can see that when you get into the DataFerrett program you can expand the Data Sets folder and expand the American Community Survey folder and you can see that you can access our 2010 through 2012 three year PUMS estimates as well as the five year multi-year PUMS estimates.

And even further down the list you can access the single year public use micro data sample as well.

So some of the other considerations to take into account when working with the PUMS is that the - if you're working with a national level file rather than individual state files we have - we break the national level file up into multiple different files that need to be concatenated.

And we have information about how to do this in the PUMS readme which once again can be accessed from our Data and Documentation page under the PUMS documentation at the link provided here on this slide.

And if you're working with both the housing and the person records and you need to merge these two files you will need to use the variable serial number that we include on the two files.

And information about how to use the serial number to merge both of these housing and personal record files together can be reviewed in the PUMS readme as well.

And then we also I mentioned earlier that the PUMS is a weighted sample so you need to use weights to create estimates. And those two weights as if that you would use would be WGTP if you're creating household level weights or PWGTP if you're interested in creating person level weights on the person's file.

And we also include PUMS replicate weights on our PUMS files. And these weights are used for calculating standard errors. And information about that is included in our accuracy, the PUMS document that I'll be talking about it a little bit.

So furthermore on estimating variances with the PUMS because PUMS is not just a simple random sample you do need to use some of the weighting that we provided, the processes that we talk about in our Accuracy of the Data or accuracy of the PUMS document to create standard errors.

And some of those ways would be to use the design factor method and or you can use the replica weights method.

And both of those methods as I was mentioning earlier can be found in our Accuracy of the PUMS document. And we detail how to go through both of those methods and how to use either or to create those standard errors.

And so some of the resources that I've been mentioning include the - are mainly accessed from our PUMS Documentation Web page at the URL listed at the top of this slide.

And on the PUMS Documentation page we include documents that lists all of the subjects that are covered in our PUMS files as well as all of the code lists.

And the code lists specifically can be used to look at the differences between some of the multi-vintages multi-vintage variables and you can see the differences between some of the variables that have multi-vintages of that are very detailed codes.

And then we also include lists of our top and bottom coded values because some of the variables in PUMS have extreme values that we need to mask to maintain confidentiality.

And you can see what those variables are as well as the top codes and the bottom codes in our top and bottom coded values document.

And then we also have the Data Dictionary which as I mentioned earlier lists out all of the different variables that are included in the PUMS files as well as descriptions of each variable and all of the values that are associated with each PUMS variable.

Another document that I've referenced quite a bit is the PUMS readme. And the readme includes a lot of information about accessing the PUMS files, how to concatenate some of the PUMS files if you're working with the national level files as well as combining the person in the housing files.

And we also included the full list of all of our multiple vintage variables in the Section 7 talking about variable changes or any new variables that we've added or variables that we've removed from a certain PUMS file.

The accuracy of the PUMS document also has a lot of useful information about working with the PUMS files about creating margins of errors and standard errors and estimate as well as a few examples of working with the

multiple vintages of the PUMS variables that are included on these 2012 three year in 2012 five year multi-year PUMS files.

And then we even have estimates for user verification that you can use to verify that you're creating PUMS estimates accurately.

And I want to specifically reference the accuracy of the PUMS document again because we do have a full section, Section 4 that deal specifically with the dual or multi-vintage variables that are now included on the 2012 multi-year PUMS files.

And so we have basic information about why we have multiple vintages as well as later on in Section 4 in the document we have two specific examples, one dealing with working with some of the detailed codes or variables with detailed codes and the multiple vintages associated with those as well as working with the multiple vintages of PUMAs.

And then I also wanted to mention a new resource that some of you might be aware of. It's the New American Community Survey Data Users group that's essentially an online community where data users can come together to discuss issues they're having with PUMS or projects they're working on and really interact directly with other data users who may be doing similar types of work or might have some novel approaches to working with the ACS data and even specifically the PUMS.

Once joined the online community you can also join a specific PUMS interest group where you can get updates and information from other data users who are specifically working with PUMS.

So this is a good resource to share with other data users some of the novel approaches to dealing with the PUMS in general but also specifically with some of these new multiple vintage or dual vintage variables and PUMAs.

The online community also lists different events that are being plan such as Webinars and conferences as well as other resources that recorded Webinars or any resources from other areas or other data users that might be useful.

And that's all that I had for today. I know I probably went pretty quickly through some of this information but I want to make sure that I left plenty of time for questions.

So I think that we'd like to have you fill out an evaluation form and then I'll be here for as many questions as any of you have.

Woman: So everyone if you could take a couple of minutes to complete the questionnaire or on the screen now for the Public Use Micro Data Sample and we will resume in about three minutes to start addressing some of the questions you might have. Thank you.

Okay everyone one more opportunity to complete the evaluation. It's really important to get your feedback on how our presenters do so please take the time to complete this and we'll resume momentarily. Thank you.

Okay everyone operator we're ready to take questions.

Coordinator: Thank you. At this time if you'd like to ask a question please pressed Star 1 on your touch-tone telephone. You will be prompted to record your name in order to be introduced.

Once again please press Star 1 on your touch-tone telephone. Please hold a moment for any questions.

Once again if you'd like to ask a question please press Star 1 on your touch-tone telephone.

Our first question will come from (Jill Wilson). Your line is open.

(Jill Williams): Hi. I have a question about using dual vintage PUMAs. The example that you looked at in Michigan was fairly simple that there was a one to one correspondence and that the geographies align completely even though the PUMA codes had changed.

But I'm wondering if you have guidance on what to do when the boundaries do indeed change? And I'm specifically looking at it from a standpoint of wanting to use data for metropolitan areas and getting as close as we can to metropolitan area boundaries using PUMAs and then just looking at how those definitions change over time makes it complicated to have two different vintages within the same file?

Tim Gilbert: Right. I know that is what I showed was one of the more straightforward examples. You can also use that MABLE GeoCorr engine through the Missouri State Data Center Web site to look where maybe now a PUMA or multiple PUMAs have been split.

And so new PUMA while it might not be a one to one you can see what PUMAs and how well they fit into other areas.

So you can get very detailed and complicated with that engine, that correspondence engine. I mean I didn't want to go into too much detail with that.

But we also do recognize that there will be instances where you just can't get an exact match. So I guess really one of I think the best tools would be to use the Missouri State data Center site to get as close as you can.

And I think you mentioned kind of talking about how the PUMAs change over time. And we have to update the PUMAs is to kind to reflect changes in population.

And we try to do that in conjunction with states and local communities as much as possible. But some of the reasons why we can't go back and kind of recreate all the PUMAs to stay as a single vintage is then we can with certain overlaps there are data disclosure concerns.

So it's not - I understand the issue there and it's not something that necessarily just has an easy quick answer.

But I would say if you've never worked with the Missouri state data centers MABLE GeoCorr that it can do a lot of detail comparisons to try to get you as close as possible.

But, you know, if you have as you're working with it if you have any further questions you can always contact us at the - either through our email or at the phone number listed here. And that goes to the branch that I work in and maybe we can get - we could also connect you up with some people from our geographic area to see if they have any more advice.

(Jill Williams): Right. Okay thank you.

Tim Gilbert: You're welcome. Thank you.

Coordinator: Our next question will come from (Jeff Rosenthal). Your line is open.

(Jeff Rosenthal): Hi. Good afternoon. I just wanted to ask kind of a clarification question or an example of using the - using the PUMS and the vintage variables for occupation.

I noticed - I was wondering if there were any additional helpful hints or examples that you can provide for using that particular variable?

Tim Gilbert: Right that's a good question. I think I meant to mention and I probably neglected to in talking about the accuracy of the PUMS document. One of the specific examples that we have in that Section 4 of that document is dealing with some of the occupation variables.

And we also in our code lists that we provide there is a crosswalk document where we talk about - that has been provided by our Industry and Occupation branch here at the Census Bureau where they give a lot of detailed information about how to work with these multiple and triple vintages.

So I would point you towards those documents. The - it does get a little more complicated than some of the dual ventures just because in this five year the 2012 five year file we do have three different vintages you kind of have to work through.

But I really think that the example we have in the Accuracy of the Data document is a good start and really kind of boiling down exactly the steps to

take in the document to look at to create the crosswalk across all three vintages.

(Jeff Rosenthal): Okay thank you.

Tim Gilbert: Thank you.

Coordinator: Before we go to our next question once again if you would like to ask a question please press Star 1 on your touch-tone telephone.

Our next question will come from (Mark Schneider). Your line is open.

(Mark Schneider): Hi. I was hoping the PUMS information might help the question I have. Our service territory is rather complex combination of county and ZIP Codes that apparently PUMS or PUMA data doesn't go down to the ZIP Code level.

Tim Gilbert: Right.

(Mark Schneider): I wonder if you have any suggestions as how to get to that combination of geographies creating them?

Tim Gilbert: Yes it is tough because like you made reference to the lowest level of the - of geography that the PUMS files cover is that PUMA level which is its fairly large at a minimum of 100,000 really.

So it is - it's not really the easiest thing to get down to small geographic areas if at all.

I would think if you're having a combinations of the counties and ZIP Codes you can using once again the Missouri State Data Centers GeoCorr and a

corresponding engine you can compare not just PUMAs to PUMAs but really a vintage of the PUMA to almost any other geographic entity so you could do ZIP Codes to PUMAs and see which ZIP Codes fit into PUMAs as well as County to PUMA or multiple different combinations of comparing geographies and how they fit together.

So I think that would probably be my best recommendation would be to try to use that tool and see how some of the geographies that you need to work with are represented through these different vintages of PUMAs that we have.

(Mark Schneider): Okay good. I appreciate it.

Tim Gilbert: Absolutely.

(Mark Schneider): We'll try...

Coordinator: Once again if you like to ask a question please press Star 1 on your touch-tone telephone. That is Star and 1. Please stand by for any further questions.

And at this time I am showing no further questions.

Tim Gilbert: I did want to mention one last thing that we should be posting these slides. I know there was a lot of documentation and a lot of different links to different Web pages.

So hopefully we'll get the slides and maybe a recording of this Webinar posted shortly so that you can have access to all of this information.

Woman: Thank you everyone for joining this Webinar today. In closing what we'll do is we get your emails in a report and I will forward the slides to you that way through email.

And I'll probably do that on Monday because I won't get the report today but I can at least get those to you the first part of next week.

Thank you for joining us today and if you need further assistance we'll leave the information for Kim on the screen for a couple of minutes but we hope that you'll join us again soon.

And if you're looking to see what else may be available through the Census Bureau check our Training Events page. We list all upcoming Webinars on that page and feel free to sign up for any that you're interested in. Thank you.

Coordinator: Thank you. That does conclude today's conference call. Thank you for participating and you may disconnect your line at this time.

END